GitHub Actions Workflow (CI/CD)

Checkout → Lint -black -flake8 → Build Docker → Unit Tests w/ pytest → Push to AWS ECR

AWS Cloud

ECR

docker run transform_data → **S3** raw data ⇄ ETL ⇄ **S3** data w/ features

DVC for version data control

training job (cloud training) → **AWS Sagemaker** → **S3** bucket model

docker run train (local training)

auto-scaling based on demand

Kubernetes w/ Kind (local cluster)

Ingress — routes to

Service (NodePort:30080)

Load Balances

Pod 1 FastAPI (port 8000)

Pod 2 FastAPI (port 8000)

Deployment (manages all pods)

HPA (auto-scales)

Docker Port Mapping:
localhost:30080 → Node:30080 → ClusterIP:80 → Pod:8000
(NodePort) (kube-proxy) (container)
used by: user other pods FastAPI

Client
http://localhost:30080
http://localhost:30080/health
http://localhost:30080/predict
http://localhost:30080/docs